Désempilement de mesures de temps de réponse par un algorithme E.M. modifié.

Tabea REBAFKA[1,2], François ROUEFF[2], Antoine SOULOUMIAC[1]

[1]CEA, LIST, 91191 Gif-sur-Yvette Cedex, France

[2]Institut Télécom, Télécom ParisTech (CNRS LTCI), 46, rue Barrault, 75634 Paris Cedex 13, France

tabea.rebafka@cea.fr, roueff@telecom-paristech.fr, antoine.souloumiac@cea.fr

Résumé – Nous proposons une méthode d'estimation rapide et efficace permettant de compenser l'effet d'empilement ("pile-up") dans des mesures de temps de réponse en fluorescence. L'empilement provient du fait que seul le plus petit temps de réponse parmi un nombre aléatoire de photons émis est détecté. Un estimateur du type maximum de vraisemblance est proposé ainsi qu'une adaptation de l'algorithme E.M. dans ce cadre inhabituel. L'algorithme E.M. proposé s'adapte particulièrement bien au contexte de mesures de fluorescence résolue en temps, pour lesquelles un modèle de mélange fini de lois exponentielles est couramment utilisé. Nous validons la méthode sur des données réelles et simulées. Comparé à la procédure couramment utilisée en fluorescence, l'algorithme E.M. modifié permet de raccourcir le temps d'acquisition environ d'un facteur 10.

Abstract – We propose a fast and efficient estimation method that compensates the pile-up effect encountered in fluorescence lifetime measurements. The pile-up effect is due to the fact that only the shortest arrival time of a random number of emitted fluorescence photons can be detected. A likelihood-based estimator is developed that can be computed by an EM-type algorithm. This modified EM algorithm is particularly well-suited for fluorescence lifetime measurements, where a mixture of exponential distributions is often used in modeling. We evaluate the method on real and simulated data. Compared to the currently used method in fluorescence, the modified EM algorithm allows a reduction of the acquisition time of about a factor 10.

1 Le problème de l'empilement

1.1 Contexte : mesures de fluorescence

Afin d'obtenir des informations sur les molécules, leur milieu (viscosité, pH, polarité) et leurs interactions avec d'autres molécules à partir d'un échantillon, les mesures de fluorescence consistent à évaluer la *durée de vie de fluorescence* des molécules, c'est-à-dire le temps entre l'excitation des molécules par une impulsion de laser et l'émission des photons de fluorescence moléculaire. Une durée de vie est de l'ordre de quelques nanosecondes et se modélise par une distribution exponentielle dont le paramètre dépend de la molécule et de son milieu. Actuellement, la technique de mesure de fluorescence la plus répandue est le TCSPC (Time Correlated Single Photon Counting), voir [3, 2, 8]. Pour des raisons techniques d'instrumentation, le détecteur de photon enregistre uniquement le premier photon reçu à la suite d'une excitation. Tout photon arrivant plus tard est perdu. Cette perte entraîne une distorsion nonlinéaire de la distribution des temps de réponses observés par rapport à la distribution initiale des durées de vie. C'est ce que l'on appelle l'effet d'*empilement* ou *pile-up*. Dans les appareils de mesure actuels, l'empilement n'est pas pris en compte et l'intensité des laser utilisés est extrêmement faible ($\lambda \leq 0,05$) pour en limiter l'importance. En effet, l'empilement n'apparaît que si plusieurs photons sont émis par excitation. Malheureusement, il en résulte des temps d'acquisition relativement long puisque pour les intensités utilisées la plupart des excitations ne conduisent à aucune émission reçue par le détecteur. Dans [6], nous avons étudié l'information associée à des mesures de ce type en fonction de l'intensité utilisée. Cette étude montre que, de ce point de vue, la valeur optimale de l'intensité est bien plus élevée que celle utilisée en pratique. Néanmoins, pour une telle intensité, l'effet d'empilement ne peut plus être négligé. En raison de la forme compliquée de la distribution des observations empilées, les estimateurs classiques comme l'estimateur du maximum de vraisemblance et l'estimateur de la méthode des moments sont infaisables. Une méthode MCMC est proposée dans [5, 6] pour résoudre numériquement le problème d'estimation des paramètres en présence de l'empilement. Cette méthode a permis de confirmer que l'augmentation de l'intensité permet effectivement une estimation plus efficace des paramètres mais présente des inconvénients pratiques, notamment un temps de calcul coûteux. L'objectif de cette contribution est de fournir une méthode facilement utilisable en pratique. Cette méthode a fait l'objet d'un dépôt de brevet [7].

1.2 Le modèle

Soit $Y_1, Y_2, \ldots$ une suite de variables aléatoires (v.a.) à valeurs dans $\mathbb{R}_+^*$, indépendantes et identiquement distribuées (i.i.d.)

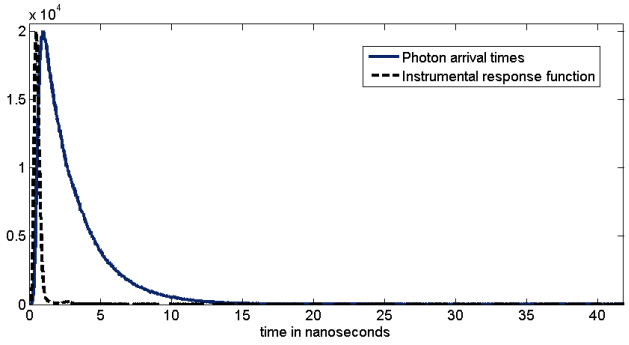

FIGURE 1 – Histogramme des mesures TCSPC et de la fonction d'appareil avec $\lambda = 0,166$.

et de fonction de répartition $F(y) = \mathbb{P}(Y_k \leq y)$. Soit N une v.a. à valeurs dans $\mathbb{N}^*$ indépendante de $(Y_k)_k$. L'observation *empilée* Z est la v.a. définie par

$$Z = \min\{Y_1, \ldots, Y_N\};.$$

Dans le contexte de la fluorescence, N est le nombre de photons de fluorescence émit à la suite d'une excitation et les Y_k représentent les temps d'arrivée de photons de fluorescence au détecteur provenant de différentes molécules.

La fonction de répartition G de Z est donnée par

$$G(z) = 1 - M(1 - F(z)) \ , \quad z \in \mathbb{R}_+ \ , \tag{1}$$

où M est la fonction génératrice des moments de N, $M(u) = \mathbb{E}[u^N]$, $u \in [0,1]$. De plus si F admet une densité f, G admet aussi une densité g donnée par

$$g(z) = f(z)\dot{M}(1 - F(z)), \quad z \in \mathbb{R}_+ \ , \tag{2}$$

où $\dot{M}(u) = \mathbb{E}[Nu^{N-1}]$, $u \in [0,1]$. Nous nous intéressons à l'estimation de F à partir d'un échantillon i.i.d. de la fonction de répartition G dans un contexte paramétrique : F est supposé admettre une densité f_θ avec θ paramètre inconnu de dimension finie. Notre objectif est d'identifier la loi initiale malgré la distorsion apportée par l'empilement. Ce problème s'apparente donc à un *problème inverse*.

Dans le cas des mesures de fluorescence, la source de photons est en général supposée poissonnienne d'intensité λ (N est alors de loi de Poisson restreinte à $\mathbb{N}^*$). La distribution initiale F est souvent un mélange de lois exponentielles de densité

$$\sum_{k=1}^{K} \alpha_k \nu_k \mathrm{e}^{-\nu_k y}, \tag{3}$$

où $\nu_k > 0$ sont de paramètres exponentiels et les proportions du mélange $(\alpha_1, \ldots, \alpha_K)^T$ vérifient $\alpha_k > 0$ et $\sum_{k=1}^K \alpha_k = 1$. On note le vecteur de paramètres par $\theta = (\alpha_1, \ldots, \alpha_K, \nu_1, \ldots, \nu_K)^T$. Dans des modèles plus réalistes, F est supposé suivre un mélange exponentiel pollué par un bruit additif.

2 Méthode de désempilement proposée

2.1 Log-vraisemblance pondérée

La maximisation de la log-vraisemblance associée aux observations du modèle d'empilement est généralement infaisable. L'idée est d'approcher la log-vraisemblance associée à un échantillon i.i.d. $(Y_1, \ldots, Y_m)$ ayant pour densité f_{θ_0}

$$L_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log(f_\theta(Y_i)) \tag{4}$$

et l'estimateur du maximum de vraisemblance correspondant :

$$\theta_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L_m(\theta) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{m} \log(f_\theta(Y_i)) \ .$$

Les propriétés du maximum de vraisemblance reposent sur le fait que $L_m(\theta)$ converge vers $\mathbb{E}_{f_{\theta_0}}[\log(f_\theta(Y))]$ qui est maximal en $\theta = \theta_0$. Nous introduirons un contraste qui converge vers la même quantité mais ce contraste sera basé sur les observations $(Z_1, \ldots, Z_n)$ ayant pour densité g_{θ_0}, définie comme la version empilée de f_{θ_0} à travers la transformation (2). La construction de ce contraste repose tout d'abord sur l'observation suivante. Soit h tel que $\mathbb{E}_{f_{\theta_0}}[h(Y)] < \infty$. Alors $\mathbb{E}_{f_{\theta_0}}[h(Y)] = \mathbb{E}_{g_{\theta_0}}\left[\frac{h(Z)}{\dot{M}(1-F_{\theta_0}(Z))}\right]$. En posant $h = \log(f_\theta)$, on obtient donc que

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\log f_\theta(Z_i)}{\dot{M}(1 - F_{\theta_0}(Z_i))} \to \mathbb{E}_{f_{\theta_0}}[\log(f_\theta(Y))] \ , \tag{5}$$

presque sûrement, quand $n \to \infty$. La suite de v.a. intervenant dans cette convergence ne peut cependant pas être utilisée pour estimer θ_0 car elle dépend de ce paramètre. On observe maintenant que M est strictement croissant de $[0,1]$ dans $[0,1]$ et on note sa réciproque M^{-1}. De plus, d'après (1), on a $1 - F_{\theta_0}(z) = M^{-1}(1 - G_{\theta_0}(z))$. Nous proposons donc de modifier (5) en remplaçant $1 - F_{\theta_0}(z)$ par $M^{-1}(1 - G_{\theta_0}(z))$ puis G_{θ_0} par $\hat{G}_n(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Z_i \leq z\}$. D'où l'approximation de $L_m(\theta)$ dans (4) donnée par

$$\begin{aligned} \hat{L}_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\log f_\theta(Z_i)}{\dot{M}(M^{-1}(1 - \hat{G}_(Z_i)))} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\log f_\theta(Z_{(i,n)})}{\dot{M}(M^{-1}(1 - \frac{i}{n}))} \ , \end{aligned}$$

où $Z_{(i,n)}$ dénote la statistique d'ordre de rang i de l'échantillon $(Z_1, \ldots, Z_n)$ qui satisfait $Z_{(1,n)} \leq Z_{(2,n)} \leq \ldots Z_{(n,n)}$. On en déduit un estimateur du paramètre θ_0 donné par

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \hat{L}_n(\theta) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{i,n} \log f_\theta(Z_{(i,n)}) \tag{6}$$

avec

$$w_{i,n} = \frac{1}{\dot{M}(M^{-1}(1 - \frac{i}{n}))}, \quad i = 1, \ldots, n \ . \tag{7}$$

TABLE 1 – Biais et écart type empirique pour chaque paramètre estimé par l'algorithm E.M. classique et modifié dans le cas où la loi initiale est un mélange exponentiel avec $K = 2$.

$\nu_1 = 0,2, \nu_2 = 2, \alpha_1 = 0,25, \alpha_2 = 0,75$								
nb d'excitations m		500		1000		5000		10 000	
E.M. classique	ν_1	0,1220	(0,3008)	0,0312	(0,1288)	0,0060	(0,0346)	0,0035	(0,0245)
$\lambda_0 = 0,05$	ν_2	1,1602	(5,6033)	0,2670	(1,2504)	0,0504	(0,2207)	0,0405	(0,1606)
	α_1	0,0613	(0,1749)	0,0145	(0,1153)	0,0014	(0,0436)	0,0031	(0,0316)
E.M. modifié	ν_1	0,0068	(0,0458)	0,0047	(0,0332)	0,0006	(0,0141)	0,0002	(0,0112)
$\lambda_0 = 2$	ν_2	0,0331	(0,2071)	0,0058	(0,1500)	0,0024	(0,0663)	0,0020	(0,0458)
	α_1	0,0045	(0,0510)	0,0023	(0,0374)	0,0005	(0,0173)	0,0002	(0,0099)

TABLE 2 – Biais et écart type empirique pour chaque paramètre estimé par l'algorithme E.M. classique et modifié dans le cas où la loi initiale est un mélange exponentiel avec $K = 3$.

$\nu_1 = 0,1, \nu_2 = 0,5, \nu_3 = 2, \alpha_1 = 0,4, \alpha_2 = 0,3, \alpha_3 = 0,3$									
nb d'excitations m		1000		5000		10 000		50 000	
E.M. classique	ν_1	0,0086	(0,0401)	0,0009	(0,0232)	0,0004	(0,0143)	0,0003	(0,0063)
$\lambda_0 = 0,05$	ν_2	0,6259	(0,8112)	0,1649	(0,3726)	0,0911	(0,2902)	0,0209	(0,1360)
	ν_3	5,5721	(32,2397)	2,1330	(10,2297)	1,6231	(9,6491)	0,1603	(0,4985)
	α_1	0,0194	(0,1658)	0,0202	(0,1149)	0,0214	(0,0900)	0,0127	(0,0403)
	α_2	0,0782	(0,1456)	0,0761	(0,1086)	0,0508	(0,0954)	0,0183	(0,0548)
	α_3	0,0977	(0,1030)	0,0560	(0,1229)	0,0293	(0,1196)	0,0056	(0,0686)
E.M. modifié	ν_1	0,0017	(0,0172)	0,0008	(0,0073)	0,0002	(0,0047)	0,0001	(0,0021)
$\lambda_0 = 1$	ν_2	0,0877	(0,2855)	0,0059	(0,1342)	0,0071	(0,0911)	0,0036	(0,0408)
	ν_3	0,8927	(3,4119)	0,0839	(0,3842)	0,0523	(0,2272)	0,0093	(0,0938)
	α_1	0,0173	(0,0970)	0,0090	(0,0469)	0,0028	(0,0265)	0,0004	(0,0156)
	α_2	0,0567	(0,0906)	0,0145	(0,0509)	0,0078	(0,0347)	0,0008	(0,0171)
	α_3	0,0393	(0,1118)	0,0055	(0,0632)	0,0051	(0,0436)	0,0012	(0,0193)

2.2 Algorithme E.M. modifié

Nous nous intéressons maintenant au problème de maximisation donné par (6). La seule différence avec celui donné par (4), hormis le fait que les observations $Y_1, \ldots, Y_m$ ont été remplacées par $Z_1, \ldots, Z_n$ est la présence des poids positifs $w_{i,n}$. Il s'en suit que si l'algorithme E.M. de [1] s'applique au problème sans empilement (4), alors ce même algorithme s'applique à la résolution du problème avec *correction d'empilement* (6). Soit S la variable latente (ou manquante) correspondant à une observation Y de densité f_θ. Notons $\pi_\theta(y, s)$ la densité jointe de (Y, S). L'algorithme E.M. modifié s'écrit de la même façon que l'algorithme E.M. si ce n'est que les poids de *correction d'empilement* (6) sont incorporés dans la partie "E" de l'algorithme. On définit une suite $(\theta^{(t)})_{t\geq 0}$ partant d'une valeur initiale $\theta^{(0)} \in \Theta$ et vérifiant la récurrence

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} w_{i,n} \mathbb{E}_{f_{\theta'}} \left[\log \pi_\theta(Z_{(i,n)}, S) | Y = Z_{(i,n)}\right] .$$

On montre facilement que la suite $(\theta^{(t)})_{t\geq 0}$ satisfait les mêmes propriétés que pour l'algorithme E.M. classique, à savoir 1) qu'à chaque itération la valeur de $\hat{L}_n(\theta^{(t)})$ augmente, 2) qu'elle converge vers un point critique de $\hat{L}_n(\theta)$.

3 Evaluation sur des données

3.1 Application à des mesures de fluorescence

Nous avons appliqué l'algorithme proposé à des données réelles de mesures de fluorescence TCSPC fournies par l'entreprise PicoQuant GmbH, Berlin, Allemagne. Pour ce type de données, il faut intégrer la *fonction d'appareil* dans le modèle de distribution des mesures (voir Figure 1). Celle-ci prend en compte les retards induits par l'appareil de mesure dans les temps acquis sous la forme d'un bruit additif.

Ces données sont obtenues avec un laser d'intensité $\lambda = 0,166$ pour laquelle environ 8% des temps acquis correspondent au minimum de 2 photons ou plus. L'empilement n'est plus négligeable dans ce cas. Le nombre de temps de réponse acquis est $n = 1\ 743\ 811$ et il y a une seule composante exponentielle, $K = 1$ dans (3). Dans cette expérience, la durée de vie est connue et donnée par $\tau = 1/\nu = 2,54$ ns. Un estimateur sans correction de l'empilement donne $\tilde{\tau} = 2,40$ ns, ce qui est trop court. Pour plus de détails sur les données voir [4].

Puisque N est supposée poissonnien de paramètre λ, les poids $w_{i,n}$ en (7) de l'estimateur de maximum de vraisemblance pondérée définie par (6) sont donnés par $w_{i,n} = \dfrac{1 - \mathrm{e}^{-\lambda}}{\lambda[\frac{i}{n}(\mathrm{e}^{-\lambda} - 1) + 1]}$.

TABLE 3 – Biais et écart type empirique pour chaque paramètre de l'estimateur par l'algorithme E.M. classique appliqué à des données empilées.

$\lambda_0 = 2, m = 10\,000$		
$\nu_1 = 0,2$	0,0579	(0,1187)
$\nu_2 = 2$	1,4268	(0,2482)
$\alpha_1 = 0,25$	0,1266	(0,0762)

Or, l'algorithme E.M. modifié donne la valeur estimée $\hat{\tau} = 1/\hat{\nu} = 2,5393$ ns. Cette expérience montre que l'estimateur proposé corrige bien l'effet d'empilement et prend correctement en compte la fonction d'appareil.

3.2 Comparaison à la méthode standard

Une comparaison de l'algorithme E.M. modifié avec une méthode utilisée typiquement en TCSPC, montre qu'une réduction importante de la variance de l'estimateur et donc du temps d'acquisition est envisageable. La performance des deux méthodes est évaluée sur des données simulées d'un modèle d'empilement où la densité initiale f_{θ_0} est la densité d'un mélange exponentiel donnée par (3) avec $K \geq 2$.

La pratique courante en fluorescence consiste à éviter tout effet d'empilement en effectuant les mesures à une très basse valeur du paramètre poissonnien λ. En effet, pour λ petit, la probabilité de plus d'un photon par excitation est négligeable, p. ex. si $\lambda = 0.05$ on a $\mathbb{P}(N > 1) \approx 0.0012$. Par conséquent, les données peuvent être considérées comme des réalisations de la loi initiale f_{θ_0}. Donc, les estimateurs standard pour des mélanges exponentiels s'appliquent. En fluorescence, la méthode des moindres carrés est fréquemment utilisée. Cependant, nous avons utilisé l'estimateur du maximum de vraisemblance classique, car contrairement à la méthode des moindres carrés, il est adapté aux petits échantillons. Donc, nous considérons comme méthode standard l'algorithme E.M. classique pour des mélanges exponentiels évalués sur des données obtenues à λ petit.

Dans cette comparaison d'estimateurs, nous voulons comparer la qualité d'estimation en fonction du temps d'acquisition. Or, en TCSPC, la distribution de N suit une loi de Poisson. Il y a donc des excitations qui ne sont pas suivies par la détection de photon, car $N = 0$. Ces excitations sont inutiles pour l'estimation des paramètres, mais elles font partie du temps d'acquisition. Dans les simulations suivantes nous avons fixé le nombre d'excitations m, qui correspond au temps d'acquisition, et puis des données pour des valeurs différentes de λ ont été simulées. En l'occurrence, on a simulé des données avec λ petit pour l'E.M. classique, afin d'obtenir des observations non empilées, et des données avec λ élevé pour l'E.M. modifié. L'échantillon pour l'E.M. classique est alors beaucoup plus petit que l'échantillon pour l'E.M. modifié, p. ex. pour $\lambda_0 = 0.05$ on a $\mathbb{P}(N = 0) = 0.951$ comparé à seulement $\mathbb{P}(N = 0) = 0.368$ lorsque $\lambda_0 = 1$. Pour des choix de paramètres différents, nous avons simulé des données du modèle d'empilement de densité g donnée par (2) et avec loi initiale multiexponentielle et N suit une loi de Poisson. Les algorithmes E.M. classique et E.M. modifié ont été appliqué avec les mêmes valeurs d'initialisation, notamment avec des poids uniformes, $\alpha_k = 1/K$, et des valeurs distinctes pour les paramètres exponentiels $\nu_1 = 1$, $\nu_2 = 3$, $\nu_3 = 5$. Les tables 1 et 2 donnent le biais et l'écart type empirique de chaque paramètre pour les deux méthodes pour différents nombres d'excitations m. Biais et écarts types empiriques sont évalués sur la base de simulation Monte Carlo comportant 1000 échantillons. Les résultats montrent que pour un temps d'acquisition fixé, c'est-à-dire un nombre d'acquisition m fixé, la qualité d'estimation obtenue par l'algorithme E.M. modifié est toujours largement supérieur à celle de l'algorithme E.M. classique. En effet, on observe qu'on obtient une qualité similaire avec seulement 10% de nombre d'excitation avec l'E.M. modifié appliqué à des données empilées. Il est donc possible de raccourcir le temps d'acquisition d'environ 90% pour obtenir des résultats d'estimation de la même qualité.

A titre d'illustration, la table 3 donne le biais et l'écart type empirique de l'algorithm E.M. sans correction, lorsque les données sont celles de la table 1 avec $\lambda = 2$ et $m = 10\,000$. L'estimateur est très biaisé et les paramètres exponentiels estimés sont trop grands, parce que l'effet d'empilement n'a pas été pris en compte.

Références

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[2] J. R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Academic/Plenum, New York, 1999.

[3] D. V. O'Connor and D. Phillips. *Time-correlated single photon counting*. Academic Press, London, 1984.

[4] M. Patting, M. Wahl, P. Kapusta, and R. Erdmann. Deadtime effects in tcspc data analysis. In *Proceedings of SPIE*, volume 6583, 2007.

[5] T. Rebafka. An MCMC approach for estimating a fluorescence lifetime with pile-up distorsion. In *GRETSI 2007 : 21ème colloque sur le traitement du signal et des images*, Troyes, France, 2007.

[6] T. Rebafka, F. Roueff, and A. Souloumiac. Information bounds and MCMC parameter estimation for the pile-up model. submitted, 2008.

[7] T. Rebafka, F. Roueff, and A. Souloumiac. Procédé d'estimation des paramètres de la distribution des temps de réponse de particules d'un système, appliqué notamment aux mesures de fluorescence. Brevet français, numéro de dépôt 09 00524, 2009.

[8] B. Valeur. *Molecular Fluorescence*. WILEY-VCH, Weinheim, 2002.